Meeting the Big Data Challenges of Climate Science through Cloud-Enabled Climate Analytics-as-a-Service

MERRA Analytic Services

John Schnase
Office of Computational and Information Sciences and Technology
NASA Goddard Space Flight Center

High-Performance Science Cloud

Dan Duffy
NASA Center for Climate Simulation
NASA Goddard Space Flight Center

“Analytics”

The discovery and communication of meaningful patterns in data.

There is extensive use of mathematics and statistics, the use of descriptive techniques, and predictive models to gain valuable knowledge from data …

… for example …

Business Analytics, Customer Analytics, Market Analytics, Fraud Analytics, Risk Analytics, Human Capital Analytics, Operations Analytics, Business Analytics, Customer Analytics, Market Analytics, Sales Analytics, Customer Services Analytics, Banking Analytics, Communications Analytics, Health Analytics, Insurance Analytics, Public Service Analytics, Retail Analytics, Learning Analytics, Web Analytics, Predictive Analytics, Prescriptive Analytics, Climate Analytics, and Analytics Analytics.

“Data Mining”

The analysis of large quantities of data to extract previously unknown interesting patterns.

There is extensive use of mathematics and statistics, the use of descriptive techniques, and predictive models to gain valuable knowledge from data …

… for example …

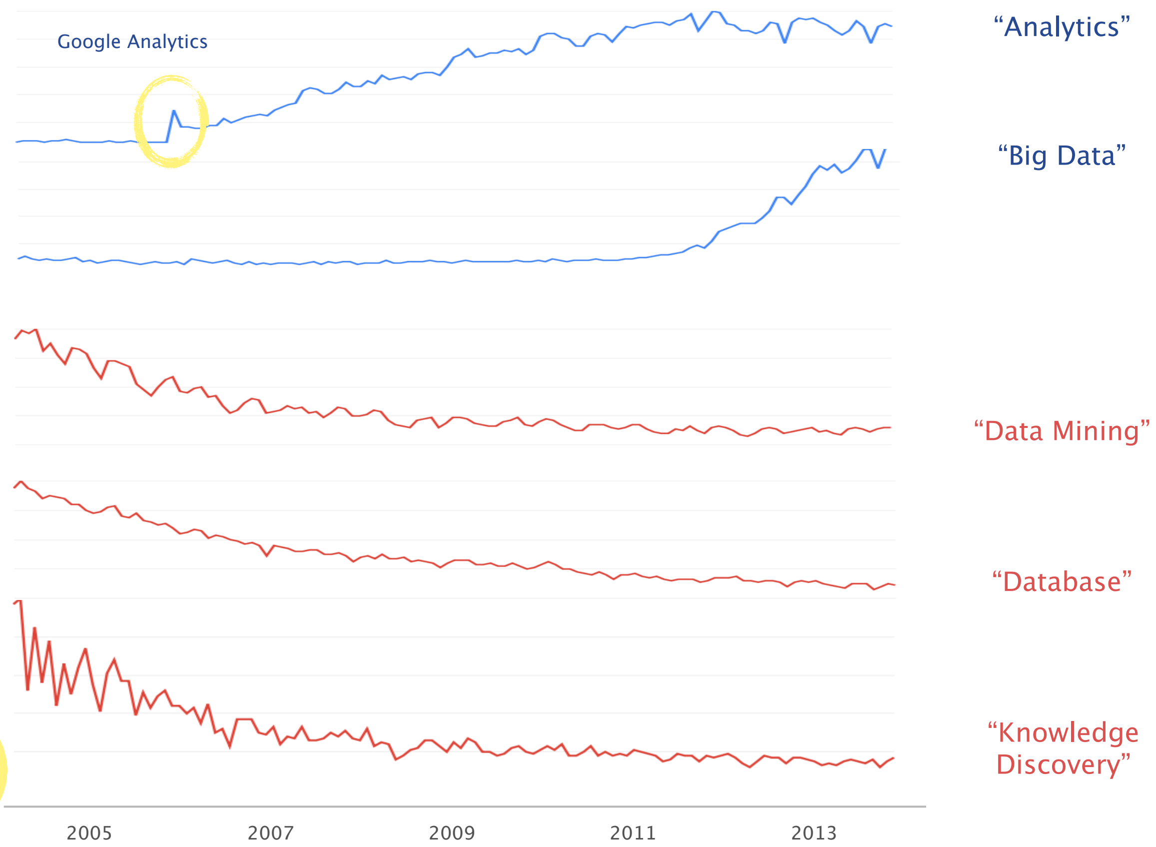
Business Data Mining, Customer Data Mining, Market Data Mining, Fraud Data Mining, Risk Data Mining, Human Capital Data Mining, Operations Data Mining, Business Data Mining, Customer Data Mining, Market Data Mining, Sales Data Mining, Customer Services Data Mining, Banking Data Mining, Communications Data Mining, Health Data Mining, Insurance Data Mining, Public Service Data Mining, Retail Data Mining, Learning Data Mining, Web Data Mining, Predictive Data Mining, Prescriptive Data Mining, Climate Data Mining, and Data Mining Data Mining.

“Analytics”

Google Trends

Hadoop, MapReduce, Cluster Computing, Big Data, Unstructured Data, Event Processing, Visualization …

Database, Artificial Intelligence, Bayesian Neural Networks, Genetic Algorithms, Machine Learning …



It’s all about scalability. At scale, even simple things become hard, even simple things become useful …

“Analytics”

We’re working on the the technology framework for climate analytics.

Right now, our analytics are simple …

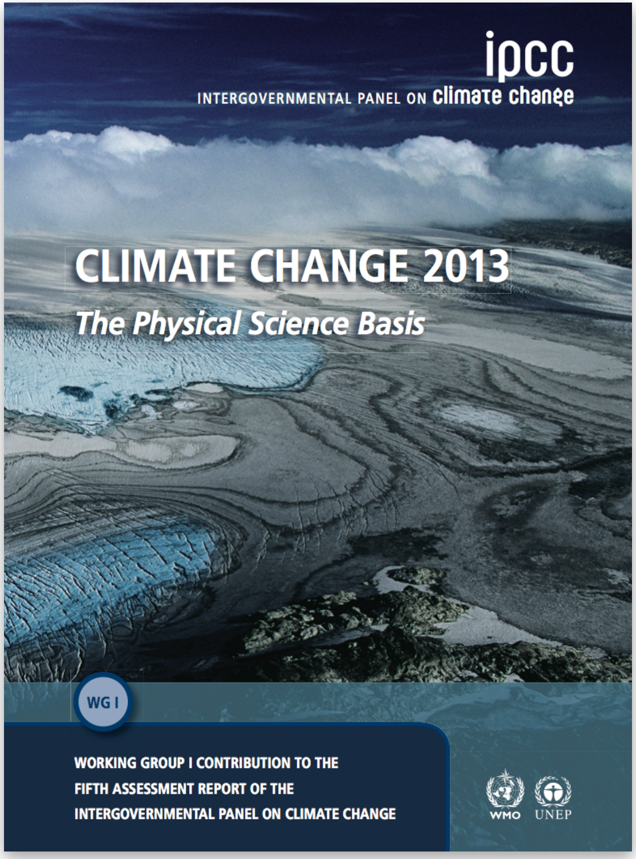
“Big Data”

Climate science is a Big Data domain.

“Big Data”

How big?

- MERRA Reanalysis Collection ~200 TB
- Total data holdings of the NASA Center for Climate Simulation (NCCS) is ~45 PB
- Intergovernmental Panel on Climate Change Fifth Assessment Report ~5 PB
- Intergovernmental Panel on Climate Change Sixth Assessment Report ~100 PB



~~"Big Data"~~

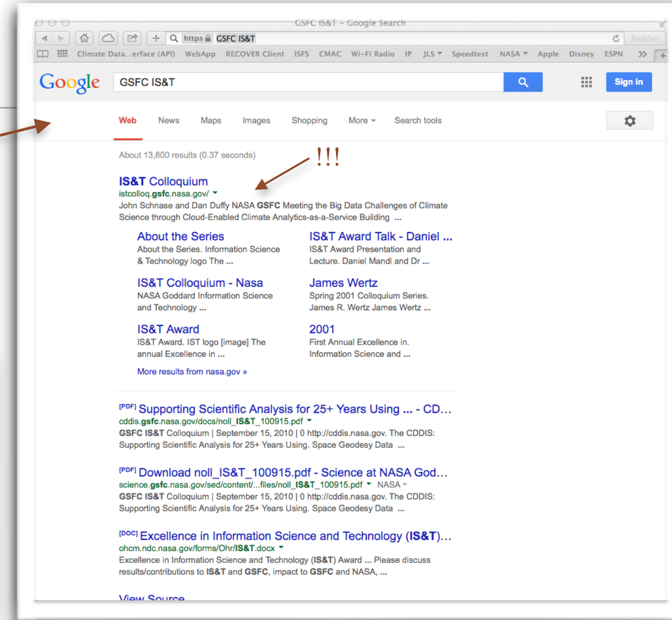
Think friction and resonance …

Data bigness depends on ease of use for the type of questions being asked …

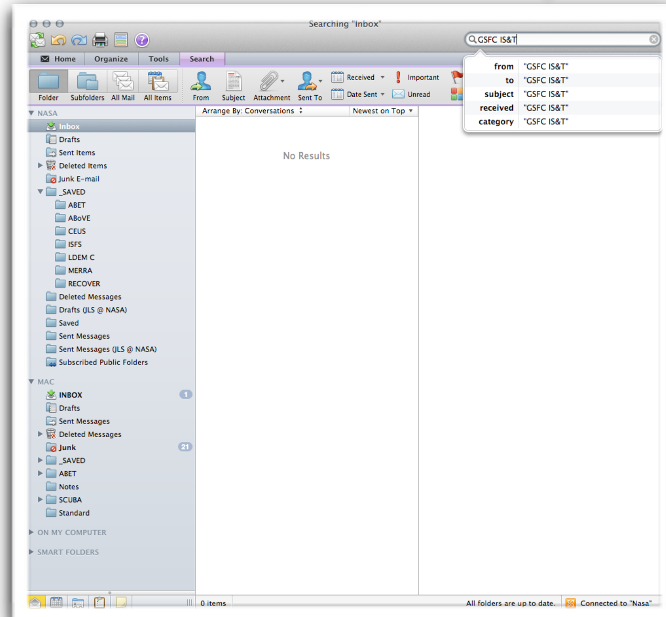
… and a particular technology may or may note help.

Query: "GSFC IS&T"

Low Friction



High Friction



Google: 13,800 results in 0.37 secs.

Outlook: No results, about as fast.

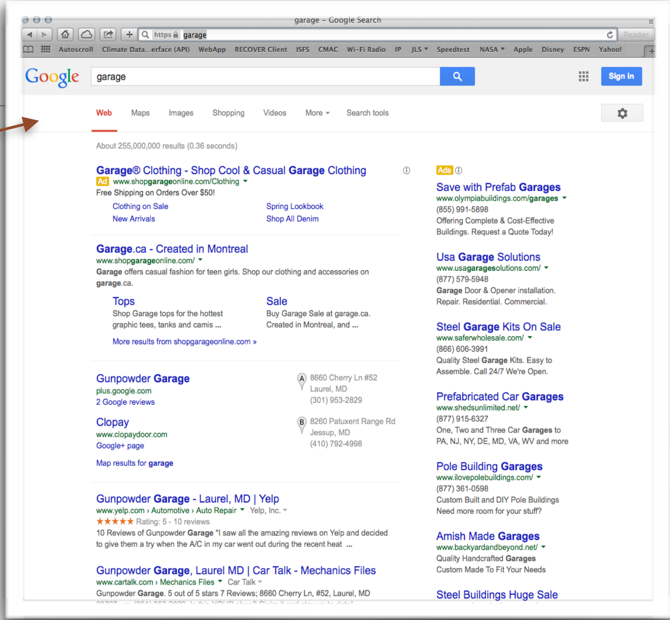
~~"Big Data"~~

Think friction and resonance …

Data bigness depends on ease of use for the type of questions being asked …

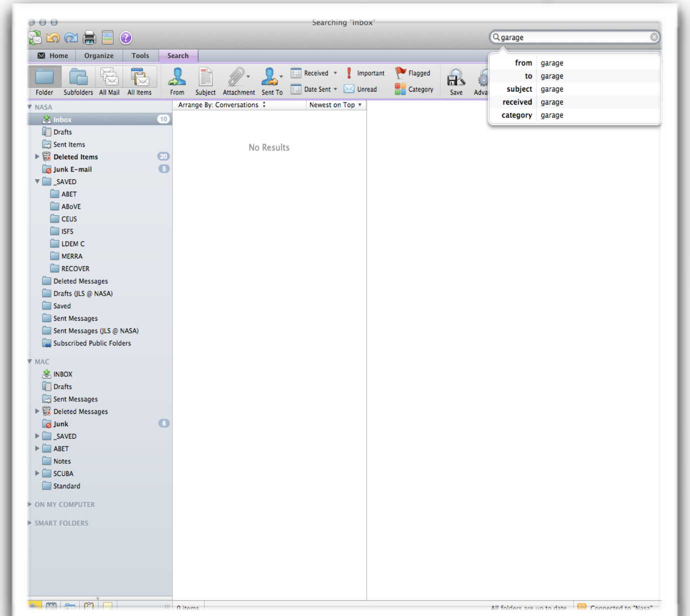
… and a particular technology may or may note help.

High Friction



High Friction

Query: "garage"



Google: 255,000,000 results in 0.36 secs.

Outlook: No results, about as fast.
(You have to select the folder to search!)

~~“Big Data”~~

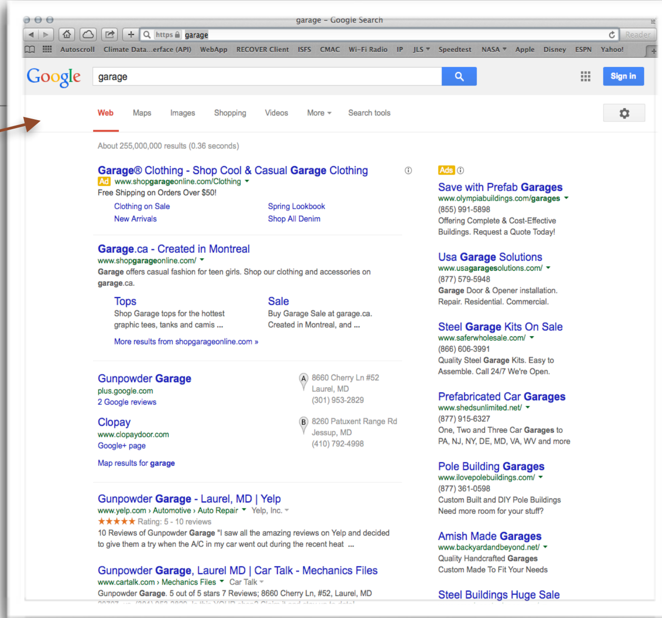
High Friction

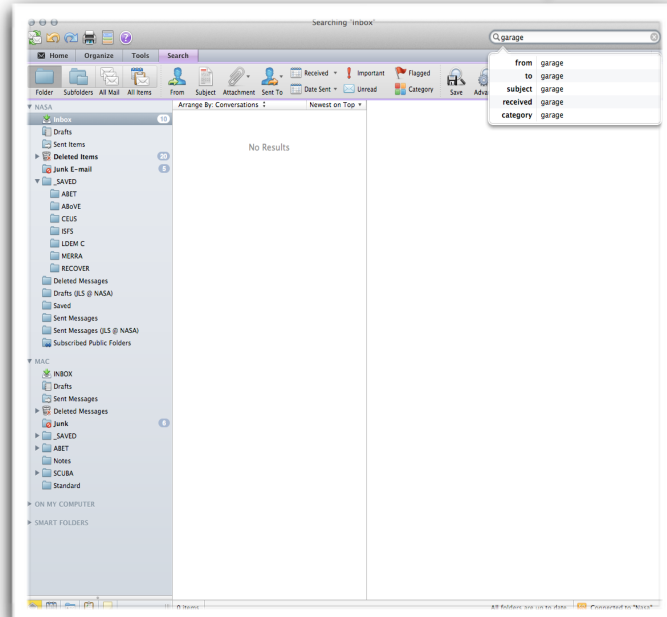
Think <u>friction</u> and <u>resonance</u> …

Data bigness depends on ease of use for the type of questions being asked …

Query: “garage”

High Friction





Successful interactions with data result when a resonance relationship sets up between data, technology, and use …

Google: 255,000,000 results in 0.36 secs.

Outlook: No results, about as fast.
(You have to select the folder to search!)

Note to Microsoft — I want to know where it is, not where it’s not …

Climate Analytics-as-a-Service

High-Performance Compute/Storage Fabric

Storage-proximal analytics
Canonical operations

Data can't move, analyses need horsepower, and leverage requires something akin to an analytical assembly language …

Data

Relevance
Collocation

Data have to be significant, sufficiently complex, and physically or logically co-located to be interesting and useful …

What are the critical resonance elements for climate analytics?

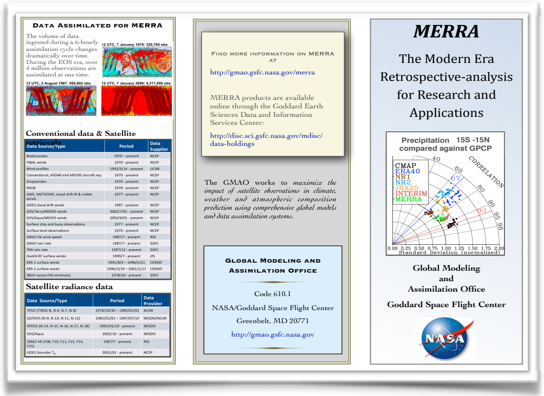
Exposure

Convenience
Extensible

Capabilities need to be easy to use and facilitate community engagement and adaptive construction …

Climate Analytics-as-a-Service

MERRA Reanalysis



Data

Relevance
Collocation

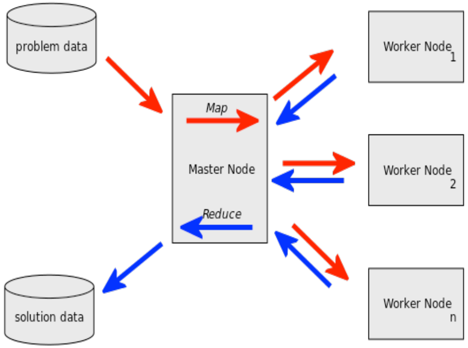
Data have to be significant, sufficiently complex, and physically or logically co-located to be interesting and useful …

High-Performance Compute/Storage Fabric

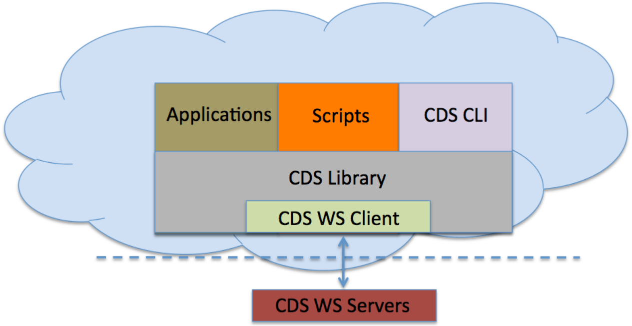
Storage-proximal analytics
Canonical operations

Data can't move, analyses need horsepower, and leverage requires something akin to an analytical assembly language …

MERRA Analytic Services



Climate Data Services API



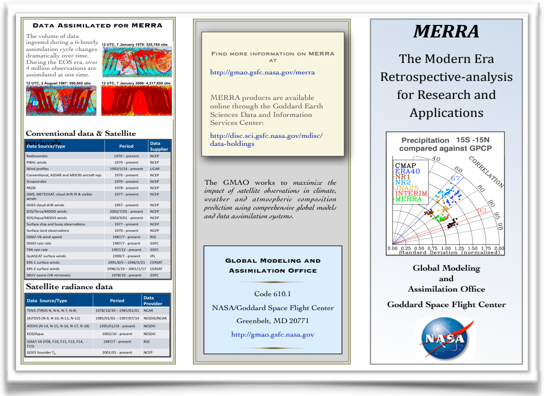
Exposure

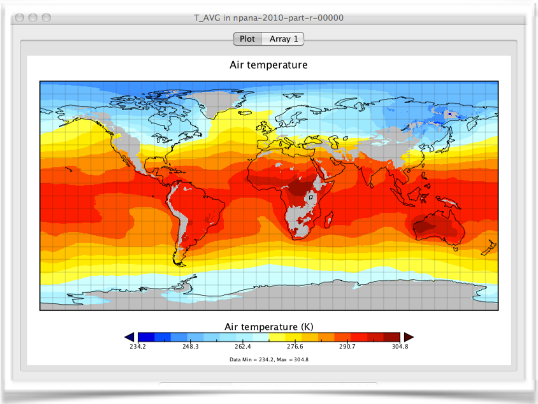
Convenience
Extensible

Capabilities need to be easy to use and facilitate community engagement and adaptive construction …

MERRA

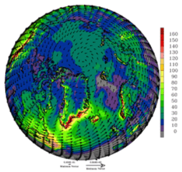
MERRA Reanalysis

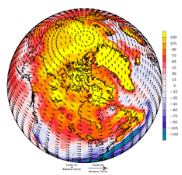




Modern Era-Retrospective Analysis for Research and Applications

- Source: Global Modeling and Assimilation Office (GMAO)
- Input: 114 observation types (land, sea, air, space) into “frozen” numerical model. (~4 million observations/day)
- Output: a global temporally and spatially consistent synthesis of 26 key climate variables. (~418 under the hood.)
- Spatial resolution: 1/2˚ latitude × 2/3˚ longitude × 42 vertical levels extending through the stratosphere.
- Temporal resolution: 6-hours for three-dimensional, full spatial resolution, extending from 1979-Present.
- ~ 200 TB, but MERRA II is on the way …





ESGF MERRA published variables

CMIP5	MERRA	Units	Description(Long Name)
rlus	rlus	W m-2	Surface Upwelling Longwave Radiation
rlut	lwtup	W m-2	TOA Outgoing Longwave Radiation
rlutcs	lwtupclr	W m-2	TOA Outgoing Clear-Sky Longwave Radiation
rsds	swgnt	W m-2	Surface Downwelling Shortwave Radiation
rsdscs	swgdnclr	W m-2	Downwelling Clear-Sky Shortwave Radiation
rsdt	swtdn	W m-2	TOA Incident Shortwave Radiation
rsut	swtdn??	W m-2	TOA Outgoing Shortwave Radiation
clt	cldtot	%	Total Cloud Fraction
pr	prectot	kg m-2 s-1	Precipitation
cl	cloud	%	Cloud Area Fraction
evspsbl	evap	kg m-2 s-1	Evaporation
hfls	eflux	W m-2	Surface Upward Latent Heat Flux
hfss	hflux	W m-2	Surface Upward Sensible Heat Flux
hur	rh	%	Relative Humidity
hus	qv	v	Specific Humidity
prc	preccon	kg m-2 s-1	Convective Precipitation
prsn	precsno	kg m-2 s-1	Snowfall Flux
prw	tqv	kg m-2	Water Vapor Path
ps	ps	Pa	Surface Air Pressure
psl	slp	Pa	Sea Level Pressure
rlds	lwgnt	W m-2	Surface Downwelling Longwave Radiation
rldscs	lwgabclr	W m-2	Surface Downwelling Clear-Sky Longwave Radiation
rsutcs	swtdn	W m-2	TOA Outgoing Clear-Sky Shortwave Radiation
ta	t	K	Air Temperature
tas	t2m	K	Near-Surface Air Temperature
tauu	taux	Pa	Surface Downward Eastward Wind Stress
tauv	tauy	Pa	Surface Downward Northward Wind Stress
tro3	o3	1.00E-09	Mole Fraction of O3
ts	ts	K	Surface Temperature
ua	u	m s-1	Eastward Wind
uas	u10m	m s-1	Eastward Near-Surface Wind
va	v	m s-1	Northward Wind
vas	v10m	m s-1	Northward Near-Surface Wind
wap	omega	Pa s-1	omega (=dp/dt)
zg	h	m	Geopotential Height

MERRA Analytic Services

MapReduce

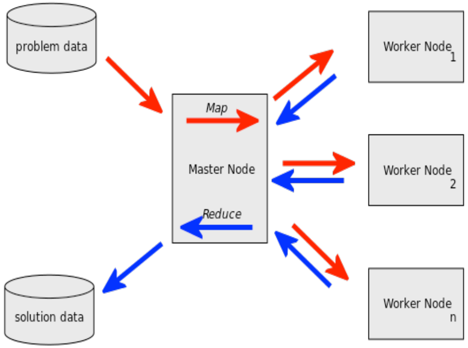
- MapReduce is a framework for processing parallelizable problems across huge datasets using a large number of computers.
- Computational processing can occur on data stored either in a filesystem (unstructured) or in a database (structured).
- MapReduce can take advantage of locality of data, processing data on or near the storage assets to decrease transmission of data.
- "Map" step: The master node takes the input, divides it into smaller sub-problems, and distributes them to worker nodes. A worker node may do this again in turn, leading to a multi-level tree structure. The worker node processes the smaller problem, and passes the answer back to its master node.
- "Reduce" step: The master node then collects the answers to all the sub-problems and combines them to form the output – the answer to the problem it was originally trying to solve.

Much of the MapReduce work has been building the code ecosystem to manage multidimensional binary NetCDF files ...

Cluster / Node Configuration

- 36 node Dell cluster, 576 Intel 2.6 GHz SandyBridge cores, 1300 TB raw storage, 1250 GB RAM, 11.7 TF theoretical peak compute capacity.
- FDR Infiniband network with peak TCP/IP speeds >20 Gbps.

MERRA Analytic Services



Canonical Ops Library

- We're also creating a small set of canonical near-storage, early-stage analytical operations that represent a common starting point in many analysis workflows in many domains. For example, avg, max, min, var, sum, count operations of the general form:

  $$result \;<=\; \text{avg}(\text{var}, (t_0,t_1), ((x_0,y_0,z_0),(x_1,y_1,z_1))),$$

  that return, in this example, the average of a variable when given a variable name, temporal extent, and spatial extent ...
- Averages over time, space, and elevation can be performed now for all MERRA variables.

Hadoop File System Organization

- Total size of the native, compressed NetCDF MERRA collection in a standard filesystem ~80 TB.
- Native MERRA files are sequenced and ingested into the Hadoop cluster in triplicated 640 MB blocks.
- Total size of MERRA/AS HDFS repository ~480 TB.

5621 lines of MapReduce code behind avg operation ...

Climate Data Services API

CDS Reference Model

Ingest – Submit/register a Submission Information Package (SIP).

Query – Retrieve data from a pre-determined service request (synchronous).

Order – Request data from a pre-determined service request (asynchronous).

Download – Retrieve a Dissemination Information Package (DIP).

Status – Track progress of service activity.

Execute – Initiate a service-definable extension. Allows for parameterized growth without API change.

CDS Library

```
Class CDSLibrary(object):

    def order(self, service, parms):
        cds_ws.order(service, parms)

    def avg(self, service, parms, destination):
        sessionId = cds_ws.order(service, parms)
        response = cds_ws.status(service, sessionId)
    ....... Loop until result is available
        cds_ws.download(service, sessionId, destination)
```

CDS CLI

```
Welcome to the NASA GSFC CISTO Climate Data Services (cds).
Type help or ? to list commands.

(nasa-gsfc-cisto-cds) order MAS parms!
GetAverageByVariable_TimeRange_SpatialExtent_VerticalExtent
&operation=avg&variable_list=T&start_date=201101&end_date=201102&a
vg_period=2&min_lon=-125&min_lat=24&max_lon=-66&max_lat=50&start
_level=13&end_level=13'

(nasa-gsfc-cisto-cds) execute HADOOP mapreduce jar!/opt/cds/bin/cds-
mas-mapreduce.jar  inputPath!/opt/cds/seq-input/merra/2011 outputPath!/
opt/cds/merra_2011_mr_seqout/npana
```

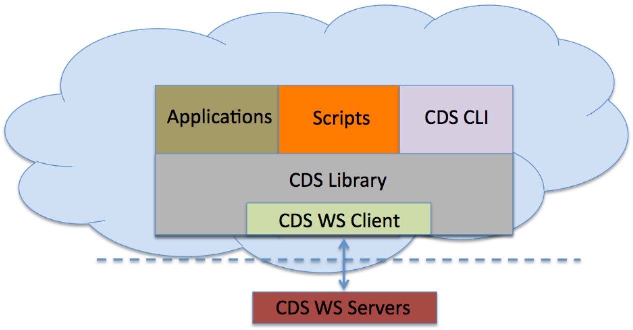
CDS Client Stack

- The MERRA/AS project has been the starting point for development of the NASA Climate Data Services (CDS) Application Programming Interface (API).
- The CDS client stack can be distributed as a software package or used to build a cloud service (SaaS) or distributable cloud image.
- This approach to API design focuses on the specific analytic requirements of the climate sciences and marries the language and abstractions of collections management (OAIS) with those of high-performance analytics (MapReduce) ...

CDS Applications

```
[gtamkin@localhost python]$ more ./user_app_ext.py
from cds import CDSApi
cds_api = CDSApi()

service = 'MAS'
north_american_parms =
'GetAverageByVariable_TimeRange_SpatialExtent_VerticalExtent
&operation=avg&variable_list=T&start_date=201101&end_date=201102&a
vg_period=2&min_lon=-125&min_lat=24&max_lon=-66&max_lat=50&start
_level=13&end_level=13'
destination=/home/gtamkin/avg-out'

Class UserAppExt(object):

if __name__ == '__main__':
    sessionId = cds_api.avg(service, north_american_parms, destination)
    print "processing complete for = " + filename
```

Climate Data Services API



CDS Scripts

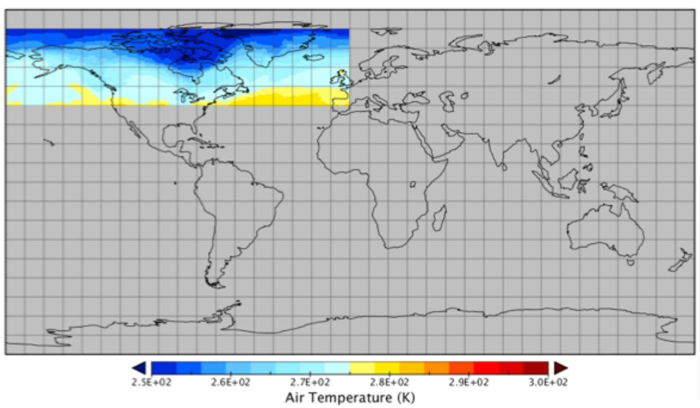
```
#!/usr/bin/env python
import time

from CDSLibrary import CDSApi
from wei_input import WEIInput
wei_exp = WEIInput()

# The rest of the file is run by the Python interpreter.
__doc__ = """This string is treated as the module docstring."""

service = wei_exp.getService()
catalog = wei_exp.getInput()
destination = wei_exp.getDestination()

cds_lib = CDSApi()
logger = cds_lib.getLogger()

start_time = time.time()

logger.debug("Generating: ca_avg_temp")
input = cds_lib.encode(catalog["ca_avg_temp_dictionary"])
cds_lib.avg(service, input, destination)

exit()
```

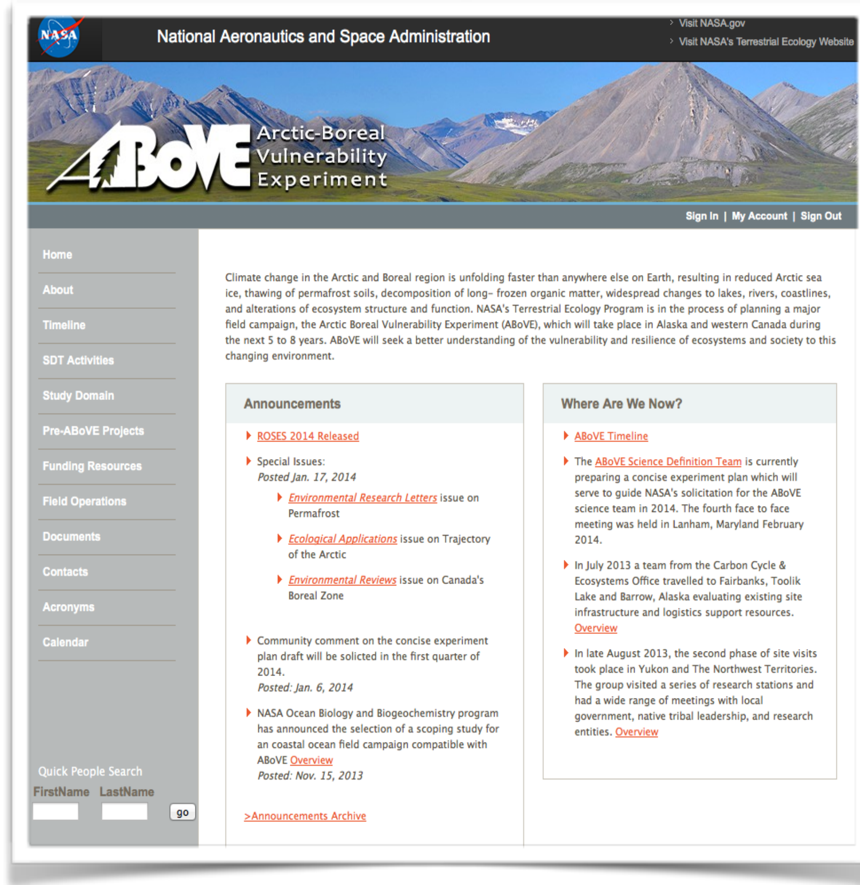
So What? Where’s the Resonance?

- Air Temperature, Precipitation / Avg, Max, Min / 1979-2014 / monthly means, 3-hourly
- Traditional: Find and order from archive (hrs?)
  Transfer ~100 GB (~1 hr, depending)
  Client-side clip/compute using GrADS
  1-1.5 days

  Server-side clipping using OPeNDAP
  (single stream op, time ??, > 2 mos)
- MERRA/AS: Server-side clip/compute (~24 hrs)
  Transfer final product ~1.5 GB

Takes about as long, but the scientist is free to work on other things ...





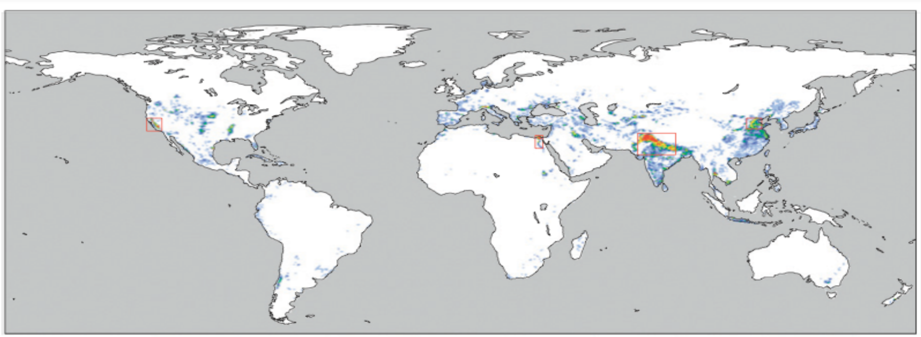
Wei Experiment

An Estimation of the Contribution of Irrigation to Precipitation Using MERRA

THE UNIVERSITY OF TEXAS AT AUSTIN

JACKSON SCHOOL OF GEOSCIENCES

- Wei team used MERRA data to study four intensively irrigated regions: northern India/Pakistan, the North China Plain, the California Central Valley, and the Nile Valley.
- Seasonal rates of evapotranspiration with and without irrigation over the studied areas were then compared to assess the impact of irrigation.
- The data required for these calculations include precipitation, evapotranspiration, temperature, humidity, and wind at different tropospheric levels at six-hourly time steps from 1979 to 2002.
- This early-stage data reduction—average values for environmental variables over specific spatiotemporal extents—is the type of data assembly that historically has been performed on the scientist's workstation after transfers from public archives of large blocks of data.



FEBRUARY 2013 WEI ET AL.

Where Does the Irrigation Water Go? An Estimate of the Contribution of Irrigation to Precipitation Using MERRA

JIANGFENG WEI*

Center for Ocean–Land–Atmosphere Studies, Calverton, Maryland

PAUL A. DIRMEYER

Department of Atmospheric, Oceanic and Earth Sciences, George Mason University, Fairfax, Virginia, and Center for Ocean–Land–Atmosphere Studies, Calverton, Maryland

DOMINIK WISSER

Department of Physical Geography, Utrecht University, Utrecht, Netherlands

MICHAEL G. BOSILOVICH

Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, Maryland

DAVID M. MOCKO

SAIC and Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, Maryland

(Manuscript received 24 May 2012, in final form 21 September 2012)

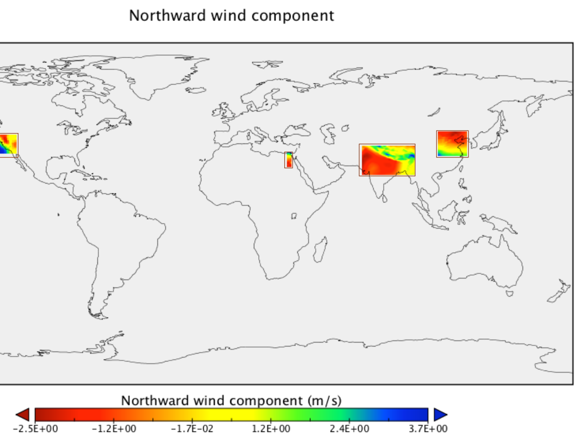
ABSTRACT

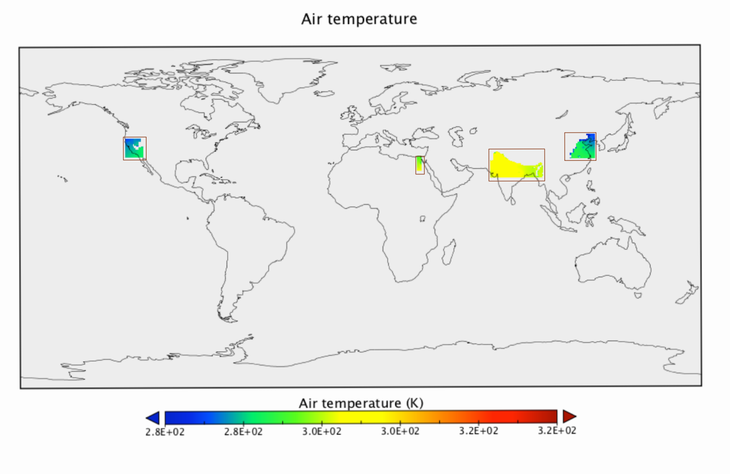
Irrigation is an important human activity that may impact local and regional climate, but current climate model simulations and data assimilation systems generally do not explicitly include it. The European Centre for Medium-Range Weather Forecasts (ECMWF) Interim Re-Analysis (ERA-Interim) shows more irrigation signal in surface evapotranspiration (ET) than the Modern-Era Retrospective Analysis for Research and Applications (MERRA) because ERA-Interim adjusts soil moisture according to the observed surface temperature and humidity while MERRA has no explicit consideration of irrigation at the surface. But, when compared with the results from a hydrological model with detailed considerations of agriculture, the ET from both reanalyses show large deficiencies in capturing the impact of irrigation. Here, a back-trajectory method is used to estimate the contribution of irrigation to precipitation over local and surrounding regions, using MERRA with observation-based corrections and added irrigation-caused ET increase from the hydrological model. Results show substantial contributions of irrigation to precipitation over heavily irrigated regions in Asia, but the precipitation increase is much less than the ET increase over most areas, indicating that irrigation could lead to water deficits over these regions. For the same increase in ET, precipitation increases are larger over wetter areas where convection is more easily triggered, but the percentage increase in precipitation is similar for different areas. There are substantial regional differences in the patterns of irrigation impact, but, for all the studied regions, the highest percentage contribution to precipitation is over local land.

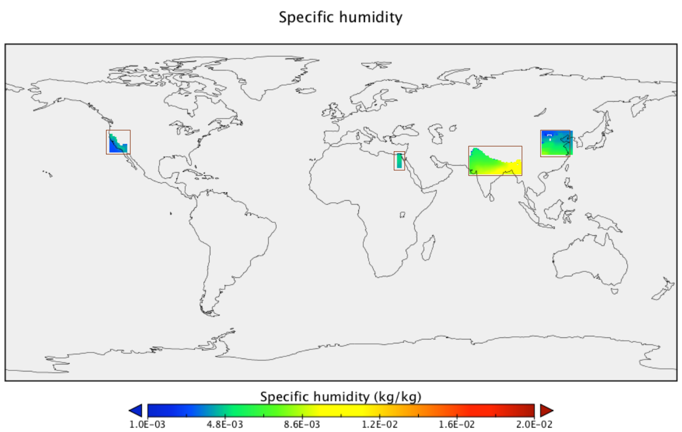
Wei, J., Dirmeyer, P. A., Wisser, D., Bosilovich, M. G., & Mocko, D. M. (2013). Where does irrigation water go? An estimate of the contribution of irrigation to precipitation using MERRA. *Journal of Hydrometeorology*, 14(2), 271–289.

Wei Experiment

An Estimation of the Contribution of Irrigation to Precipitation Using MERRA







Wei, et al.

- ~8.4 TB transferred from archive to local workstation (weeks)
- Clipping, averaging performed by Fortran program on local workstation (days)

MERRA/AS (Time trials in progress ...)

- Clipping, averaging performed by MERRA/AS (~28 hrs)
- ~35 GB of final product moved to local workstation

- Significant time savings in data wrangling,

- rapid screening over monthly means files takes minutes, and

- there's a possibility of folding Dr. Wei's modeling algorithm back into the CDS API ...

RECOVER

Rehabilitation Capability Convergence for Ecosystem Recovery

An Automated Burned Area Emergency Response Decision Support System
for Post-fire Rehabilitation Management of Savanna Ecosystems in the Western US

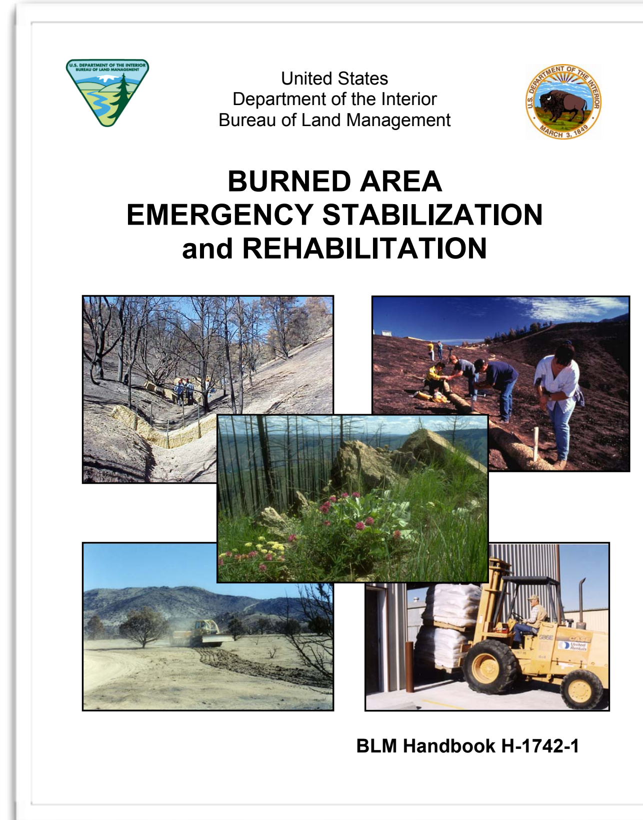
Keith T. Weber
GIS Training and Research Center
Idaho State University

John L. Schnase[1], Molly E. Brown[2], and Mark Carroll[2]
[1]Office of Computational and Information Sciences and Technology
[2]Biospheric Science Branch
NASA Godard Space Flight Center

RECOVER

- After a major wildfire, law requires that the federal land management agencies certify a comprehensive plan for public safety, burned area stabilization, resource protection, and site recovery.
- These BAER plans are due within 14 days of containment of a major wildfire and become the guiding document for managing the activities and budgets for all subsequent remediation efforts.
- Post-fire rehabilitation planning is a data-intensive process and requires better access to new types of data products …

e.g MERRA, SMAP, …


United States
Department of the Interior
Bureau of Land Management

BURNED AREA
EMERGENCY STABILIZATION
and REHABILITATION

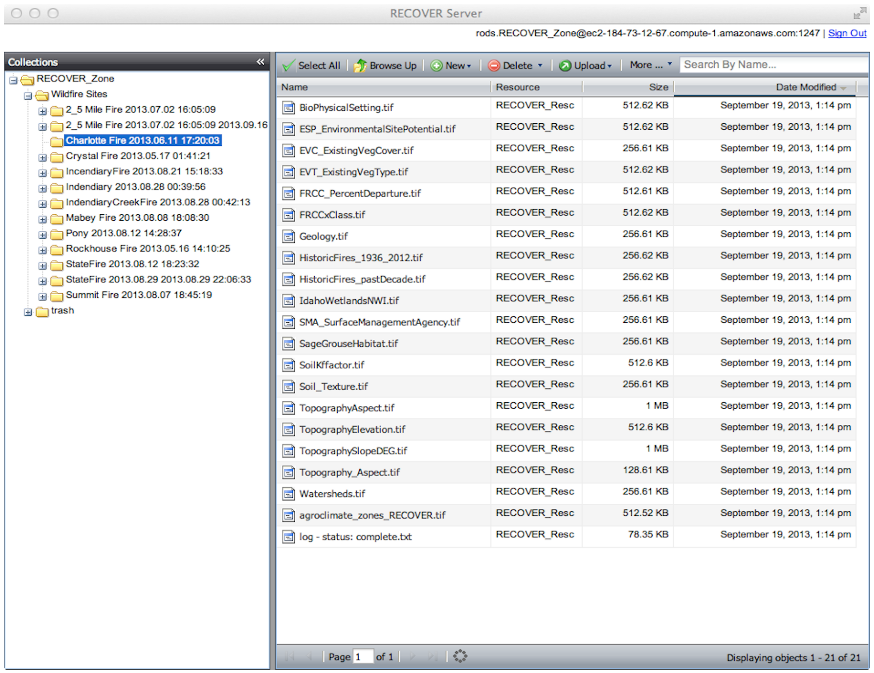
BLM Handbook H-1742-1

RECOVER

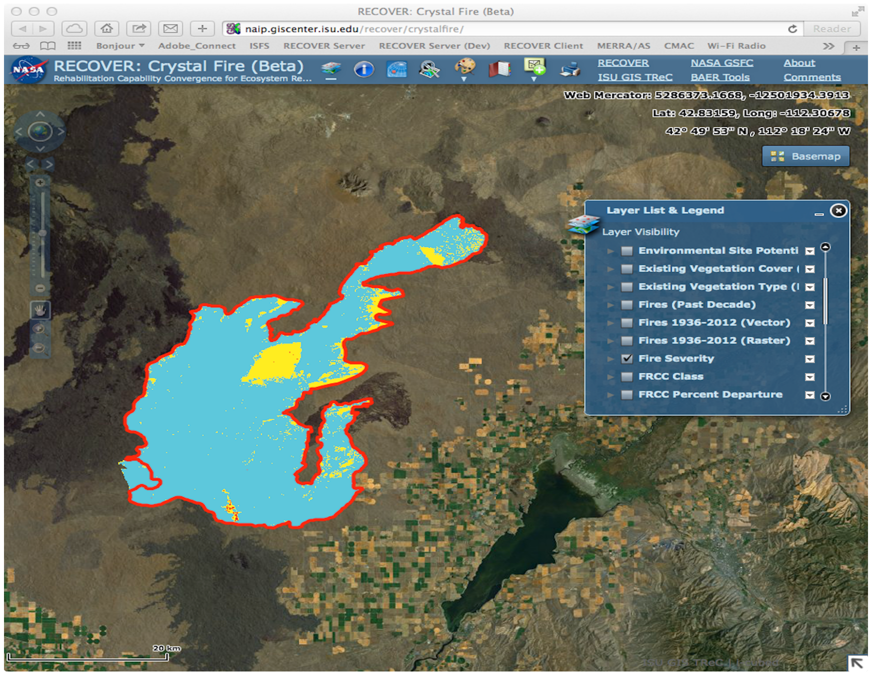
- RECOVER is a site-specific decision support system bringing together all the information necessary for post-fire rehabilitation decision-making.
- Designed in close collaboration with the US Department of Interior Bureau of Land Management (BLM) and Idaho Department of Lands (IDL).
- Uses rapid resource allocation capabilities of cloud computing to automatically gather data from various web services.
    - Earth observational data
    - Derived decision products
    - Historic biophysical layers
- <u>Automated data assembly</u> provides operational partners a complete and ready-to-use analysis environment customized for target wildfires.
- RECOVER is transforming this information-intensive process by <u>reducing from days to a matter of minutes</u> the time required to assemble and deliver crucial wildfire-related data.

RECOVER

RECOVER Server



RECOVER Client



For YouTube demonstrations, please see:

http://www.youtube.com/watch?v=LQKi3Ac7yNU	RECOVER Server
http://www.youtube.com/watch?v=SGhPpiSYpVE	RECOVER Client

GIS Training and Research Center

RECOVER

- More than a dozen agency collaborators participated in the Phase 1 feasibility study.
- The system was used in Idaho in six actual fires in the 2013 fire season.
- More than two dozen data layers assembled on average in 60 minutes.
  ~ 90 sec. to automatically gather 20+ layers
  ~ 60 min. to manually assemble the remaining specialized, site-specific layers

Fire	Start Date	End Date	Acres Burned	RECOVER Response Time (min)	RECOVER Client URL
Crystal	15-Aug-06	31-Aug-06	220,000	N/A	http://naip.giscenter.isu.edu/recover/CrystalFire
Charlotte	2-Jul-12	10-Jul-12	1,029	N/A	http://naip.giscenter.isu.edu/recover/CharlotteFire
2 ½ Mile	2-Jul-13	3-Jul-13	924	30	http://naip.giscenter.isu.edu/recover/2nHalfMileFire
Mabey	8-Aug-13	19-Aug-13	1,142	120	http://naip.giscenter.isu.edu/recover/MabeyFire
Pony	11-Aug-13	27-Aug-13	148,170	35	http://naip.giscenter.isu.edu/recover/PonyFire
State Line	12-Aug-13	18-Aug-13	30,206	40	http://naip.giscenter.isu.edu/recover/StateFire
Incendiary Creek	18-Aug-13	n/a	1,100	90	http://naip.giscenter.isu.edu/recover/IncendiaryFire
Ridgetop	28-Jul-12	n/a	16,616	4	http://naip.giscenter.isu.edu/recover/Ridgetop_v2fire/

Late Breaking News

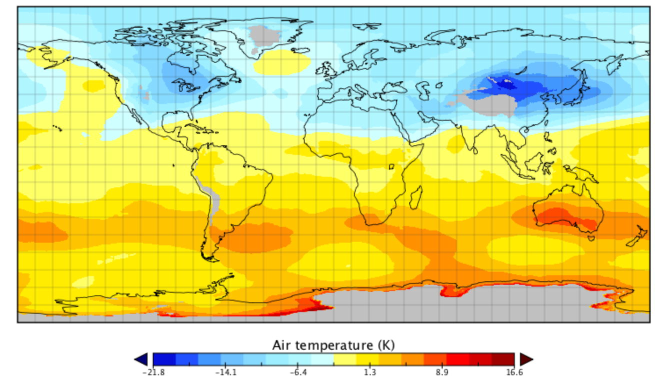
Nadeau's Standardized Temperature Anomaly …

- Period: 1 month
- Collection: instM_3d_ana_Np
- Time span: January — December 2011
- Coverage: Global
- Levels: 1 — 42 (0.1 hPa — 1000 hPa)

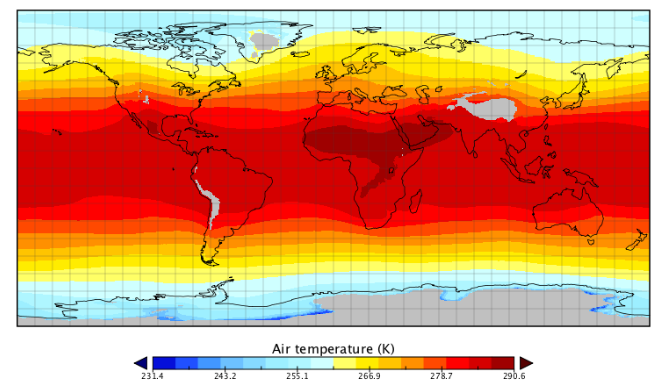
- Traditional: Find and order from archive (hrs?)
  Transfer ~10 GB (~15 min, depending)
  Client-side clip/compute using GrADS
  1-1.5 days
- MERRA/AS: One line in a python script *
  3 minute run time
  Final product ~0.5 GB

** Will be added to CDS Library …*

January



Average Global Air Temperature



Climate Analytics-as-a-Service

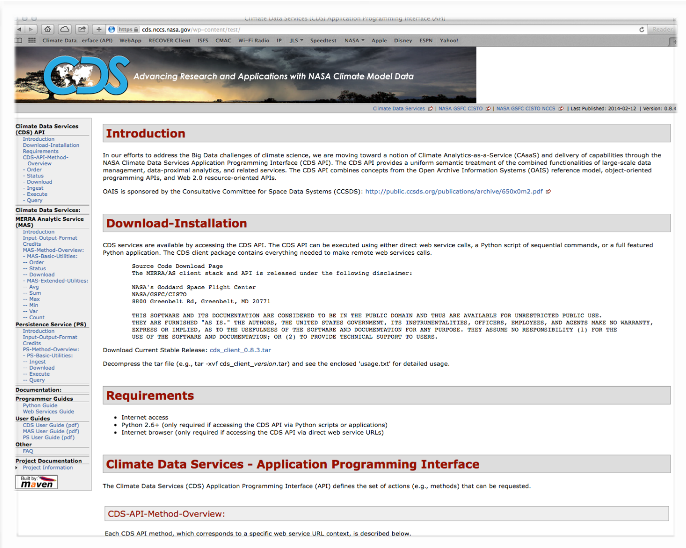
Who's interested?

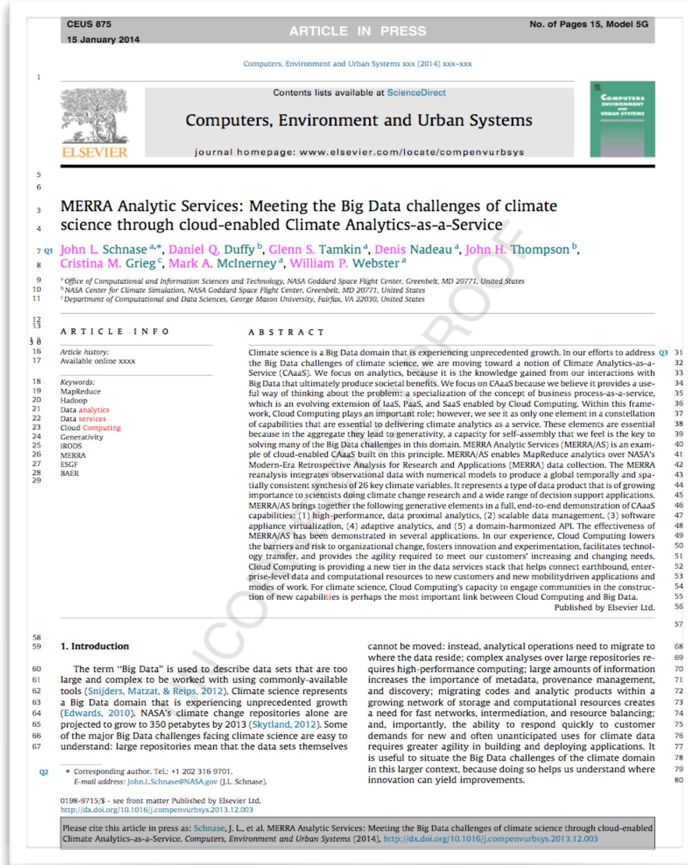
- Energy
- Education
- Agriculture
- Climate analytics
- Insurance industry
- Department of Interior
- The White House (climate.gov)

Climate Analytics-as-a-Service

Next steps

- Beta testing, add other reanalyses
- Operational deployment via Climate Data Services …





CEUS 875 — ARTICLE IN PRESS — No. of Pages 15, Model 5G

15 January 2014

Computers, Environment and Urban Systems xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/compenvurbsys

ELSEVIER

MERRA Analytic Services: Meeting the Big Data challenges of climate science through cloud-enabled Climate Analytics-as-a-Service

John L. Schnase [a,*], Daniel Q. Duffy [b], Glenn S. Tamkin [a], Denis Nadeau [a], John H. Thompson [b], Cristina M. Grieg [c], Mark A. McInerney [a], William P. Webster [a]

[a] Office of Computational and Information Sciences and Technology, NASA Goddard Space Flight Center, Greenbelt, MD 20771, United States
[b] NASA Center for Climate Simulation, NASA Goddard Space Flight Center, Greenbelt, MD 20771, United States
[c] Department of Computational and Data Sciences, George Mason University, Fairfax, VA 22030, United States

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
MapReduce
Hadoop
Data analytics
Data services
Cloud Computing
Generativity
iRODS
MERRA
ESGF
BAER

ABSTRACT

Climate science is a Big Data domain that is experiencing unprecedented growth. In our efforts to address the Big Data challenges of climate science, we are moving toward a notion of Climate Analytics-as-a-Service (CAaaS). We focus on analytics, because it is the knowledge gained from our interactions with Big Data that ultimately produce societal benefits. We focus on CAaaS because we believe it provides a useful way of thinking about the problem: a specialization of the concept of business process-as-a-service, which is an evolving extension of IaaS, PaaS, and SaaS enabled by Cloud Computing. Within this framework, Cloud Computing plays an important role; however, we see it as only one element in a constellation of capabilities that are essential to delivering climate analytics as a service. These elements are essential because in the aggregate they lead to generativity, a capacity for self-assembly that we feel is the key to solving many of the Big Data challenges in this domain. MERRA Analytic Services (MERRA/AS) is an example of cloud-enabled CAaaS built on this principle. MERRA/AS enables MapReduce analytics over NASA's Modern-Era Retrospective Analysis for Research and Applications (MERRA) data collection. The MERRA reanalysis integrates observational data with numerical models to produce a global temporally and spatially consistent synthesis of 26 key climate variables. It represents a type of data product that is of growing importance to scientists doing climate change research and a wide range of decision support applications. MERRA/AS brings together the following generative elements in a full, end-to-end demonstration of CAaaS capabilities: (1) high-performance, data proximal analytics, (2) scalable data management, (3) software appliance virtualization, (4) adaptive analytics, and (5) a domain-harmonized API. The effectiveness of MERRA/AS has been demonstrated in several applications. In our experience, Cloud Computing lowers the barriers and risk to organizational change, fosters innovation and experimentation, facilitates technology transfer, and provides the agility required to meet our customers' increasing and changing needs. Cloud Computing is providing a new tier in the data services stack that helps connect earthbound, enterprise-level data and computational resources to new customers and new mobilitydriven applications and modes of work. For climate science, Cloud Computing's capacity to engage communities in the construction of new capabilities is perhaps the most important link between Cloud Computing and Big Data.

Published by Elsevier Ltd.

1. Introduction

The term "Big Data" is used to describe data sets that are too large and complex to be worked with using commonly-available tools (Snijders, Matzat, & Reips, 2012). Climate science represents a Big Data domain that is experiencing unprecedented growth (Edwards, 2010). NASA's climate change repositories alone are projected to grow to 350 petabytes by 2013 (Skytland, 2012). Some of the major Big Data challenges facing climate science are easy to understand: large repositories mean that the data sets themselves cannot be moved: instead, analytical operations need to migrate to where the data reside; complex analyses over large repositories requires high-performance computing; large amounts of information increases the importance of metadata, provenance management, and discovery; migrating codes and analytic products within a growing network of storage and computational resources creates a need for fast networks, intermediation, and resource balancing; and, importantly, the ability to respond quickly to customer demands for new and often unanticipated uses for climate data requires greater agility in building and deploying applications. It is useful to situate the Big Data challenges of the climate domain in this larger context, because doing so helps us understand where innovation can yield improvements.

* Corresponding author. Tel.: +1 202 316 9701.
E-mail address: John.L.Schnase@NASA.gov (J.L. Schnase).

0198-9715/$ - see front matter Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.compenvurbsys.2013.12.003

Please cite this article in press as: Schnase, J. L., et al. MERRA Analytic Services: Meeting the Big Data challenges of climate science through cloud-enabled Climate Analytics-as-a-Service. *Computers, Environment and Urban Systems* (2014), http://dx.doi.org/10.1016/j.compenvurbsys.2013.12.003

Meeting the Big Data Challenges of Climate Science through Cloud-Enabled Climate Analytics-as-a-Service

MERRA Analytic Services

John Schnase
Office of Computational and Information Sciences and Technology
NASA Goddard Space Flight Center

High-Performance Science Cloud

Dan Duffy
NASA Center for Climate Simulation
NASA Goddard Space Flight Center